

Entropic selection of concepts unveils hidden topics in documents corpora

Alessio Cardillo

Department of Engineering Mathematics, University of Bristol, Bristol (UK)

<http://bifi.es/~cardillo/>

Wednesday 19th December 2018,
Workshop: Perspectives on Complex Systems 2018,
Technische Universität Berlin, Germany



Flood of information

The screenshot shows a news blog post from the **nature** news blog. The header includes the text "Nature brings you breaking news from the world of science". Below the header, the breadcrumb navigation shows "News & Comment > News blog Archive > Post". The main content features a link to a previous post about climate change and another link to a next post about animal rights. The title of the post is "Global scientific output doubles every nine years", dated 07 May 2014 at 16:46 GMT, and posted by Richard Van Noorden under the category "Policy, Publishing".

newsblog
Nature brings you breaking news from the world of science

News & Comment > News blog Archive > Post

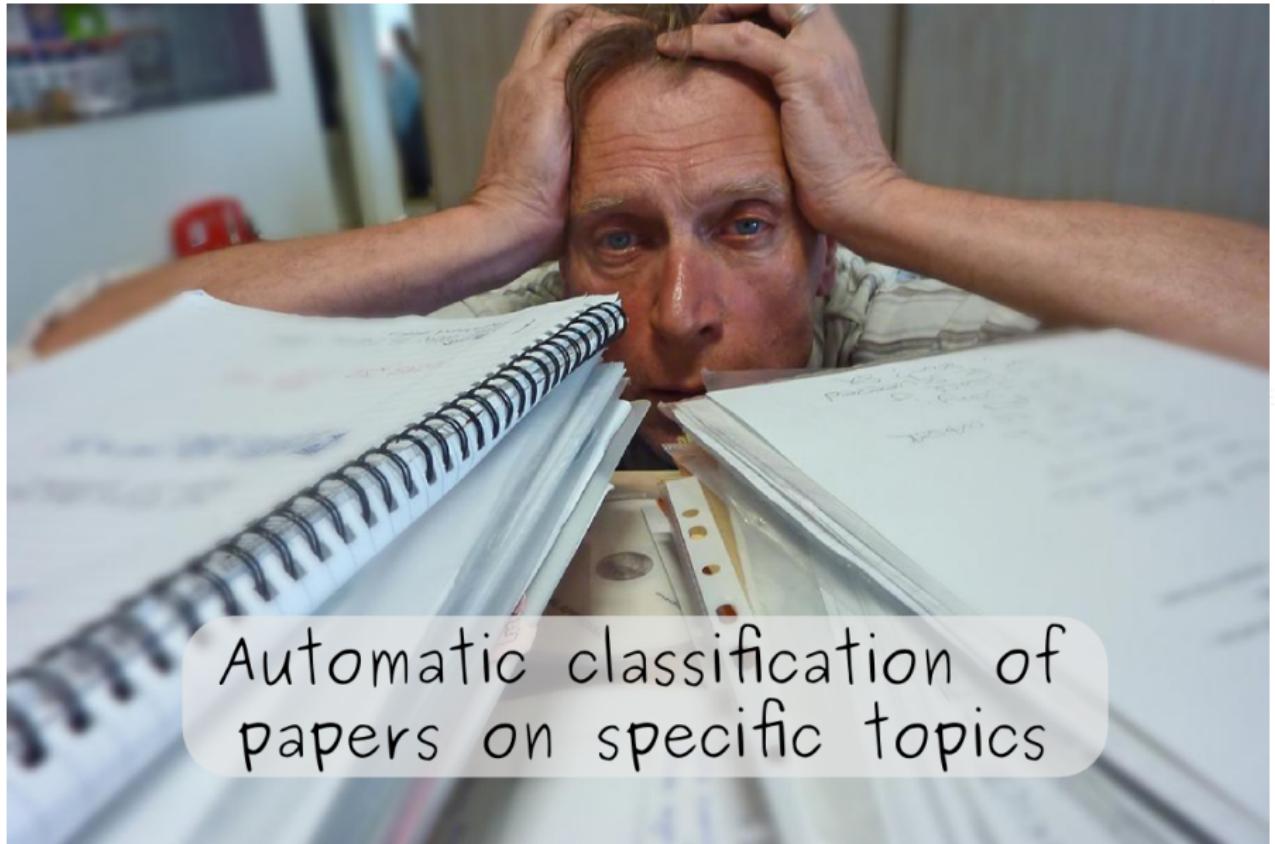
Previous post [Climate change is present danger, US warns](#) Next post [German research agencies condemn animal-rights attack on neuroscientist](#)

NEWS BLOG

Global scientific output doubles every nine years

07 May 2014 | 16:46 GMT | Posted by Richard Van Noorden | Category: Policy, Publishing

Flood of information



Automatic classification of
papers on specific topics

Flood of information

The screenshot shows a web page from the **nature** website, which is described as an "International weekly journal of science". The page has a dark red header bar with white text. Below the header, there is a navigation menu with links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and Forum. A breadcrumb navigation path is visible, showing the article's location: Archive > Volume 513 > Issue 7516 > Toolbox > Article. The main title of the article is "How to tame the flood of literature". Below the title, a subtitle reads: "Recommendation services claim to help researchers keep up with the most important papers without becoming overwhelmed." The author's name, Elizabeth Gibney, is listed, along with the publication date, 03 September 2014. There are also share and print icons at the top right of the article area.

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | Forum

Archive > Volume 513 > Issue 7516 > Toolbox > Article

NATURE | TOOLBOX

KUAEWm

How to tame the flood of literature

Recommendation services claim to help researchers keep up with the most important papers without becoming overwhelmed.

Elizabeth Gibney

03 September 2014

What do we need?

*There is an inherent problem to giving you information that you weren't actively searching for. **It has to be relevant** – so that we are not wasting your time – **but not too relevant**, because you already know about those articles.*

Anurag Acharya
Google Scholar co-creator

*Semantic Scholar offers a few innovative features, including picking out the **most important keywords and phrases** from the text without relying on an author or publisher to key them in. “**It’s surprisingly difficult for a system to do this,**”*

Oren Etzioni
CEO of AI2 (Semantic Scholar)

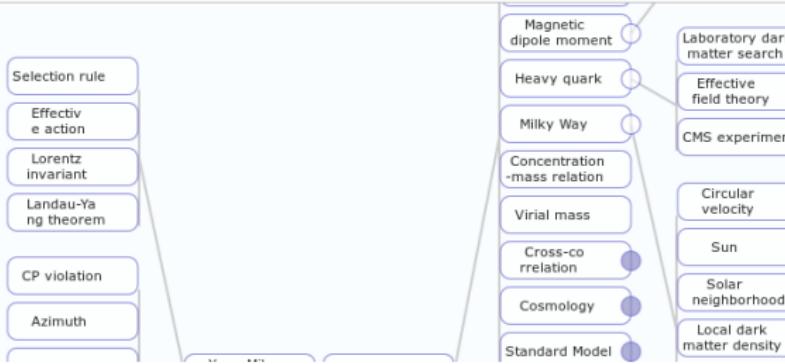
What do we need?

ScienceWISE [Ontology](#) [Bookmarks](#) [New articles](#) [News](#) [Introduction](#) [Logout](#)

Physics [Life Sciences beta](#) [Digital Humanities](#) [Information Technologies](#)

Search 

Recent ontology graph



Recently bookmarked papers

Properties of a possible class of particles ...
[astro-ph/9505117 Luis Gonzalez-Mestres](#)

The apparent Lorentz invariance of the laws of physics
 ...

Introduction to the Standard Model and E ...
[0901.0241 Paul Langacker](#)

A concise introduction is given to the standard model. Including the structure of the QCD and electroweak Lagrangians, spontaneous symmetry breaking, experimental tests, and problems.

[Standard Model](#) [Quantum chromodynamics](#) [Weak interaction](#) ...

<http://sciencewise.info>

Topic Modeling & LDA in a nutshell

PHYSICAL REVIEW X

Highlights Recent Subjects Accepted Collections Authors Referees Search Press

Open Access

High-Reproducibility and High-Accuracy Method for Automated Topic Classification

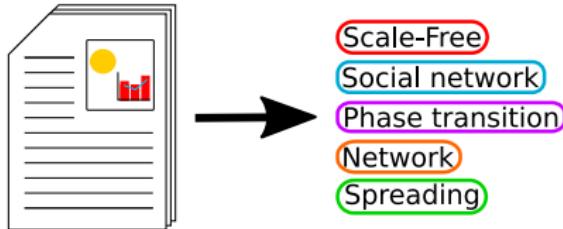
Andrea Lancichinetti, M. Irmak Sirer, Jane X. Wang, Daniel Acuna, Konrad Körding, and Luís A. Nunes Amaral
Phys. Rev. X 5, 011007 – Published 29 January 2015

- A. Lancichinetti *et al.* "High-Reproducibility and High-Accuracy Method for Automated Topic Classification". Phys.

Rev. X 5 011007 (2015)

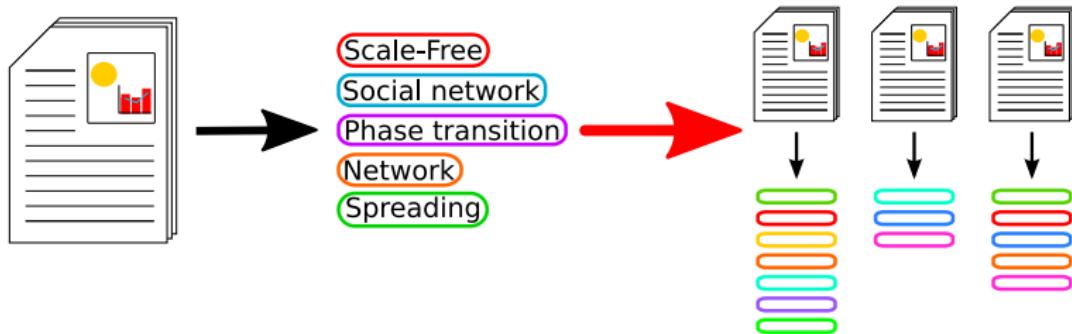


Topic Modeling & LDA in a nutshell



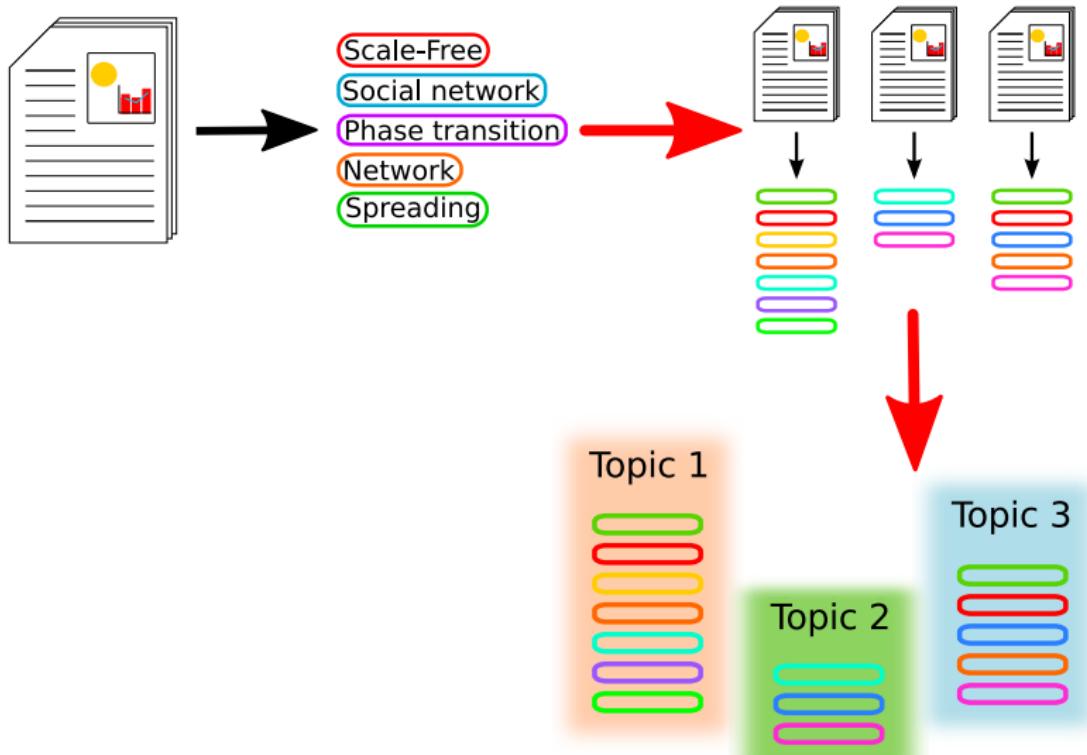
KUKUAEWm

Topic Modeling & LDA in a nutshell

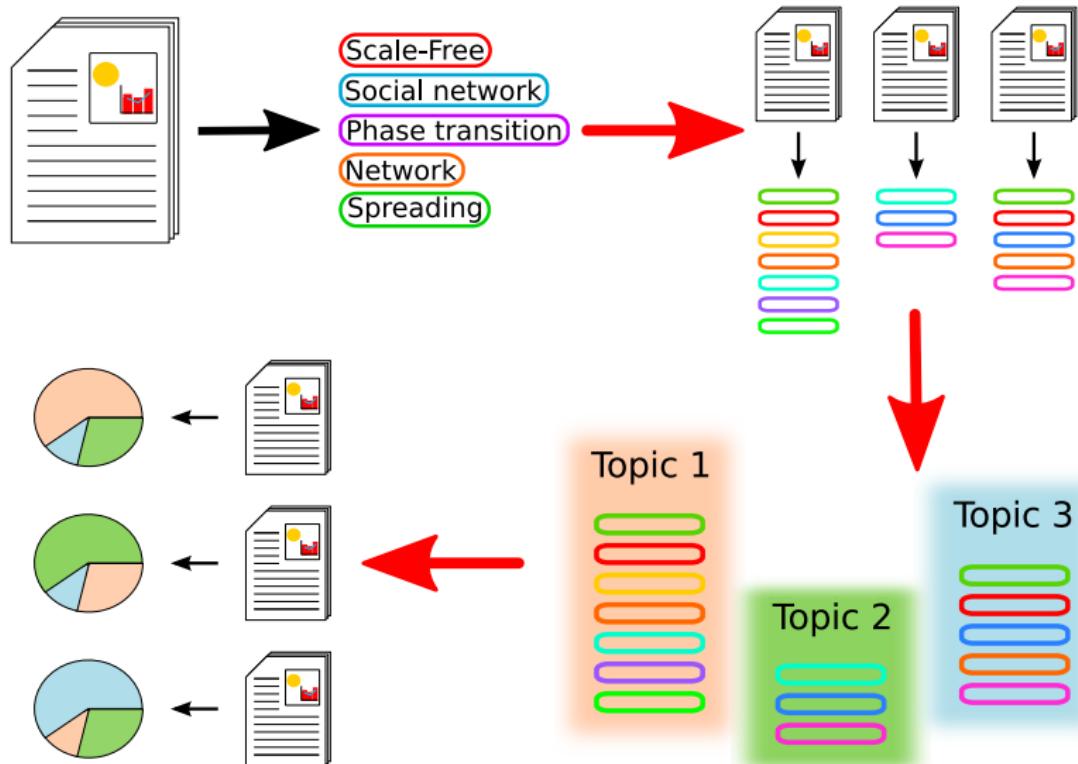


KUKUAEWm

Topic Modeling & LDA in a nutshell



Topic Modeling & LDA in a nutshell



Topic Modeling & LDA in a nutshell

KUKUAEWm

Topic Modeling & LDA in a nutshell



KUAEWm

Topic Modeling & LDA in a nutshell



Problem

Not all the words/concepts are equally **relevant** to determine the topic of a document!



Topic Modeling & LDA in a nutshell

BRACE YOURSELVES



FILTERING IS COMING

Relevant concepts



KUKUAEWm

Relevant concepts

Key features

- # of documents a concept appears in

$df_c \rightarrow$ document frequency

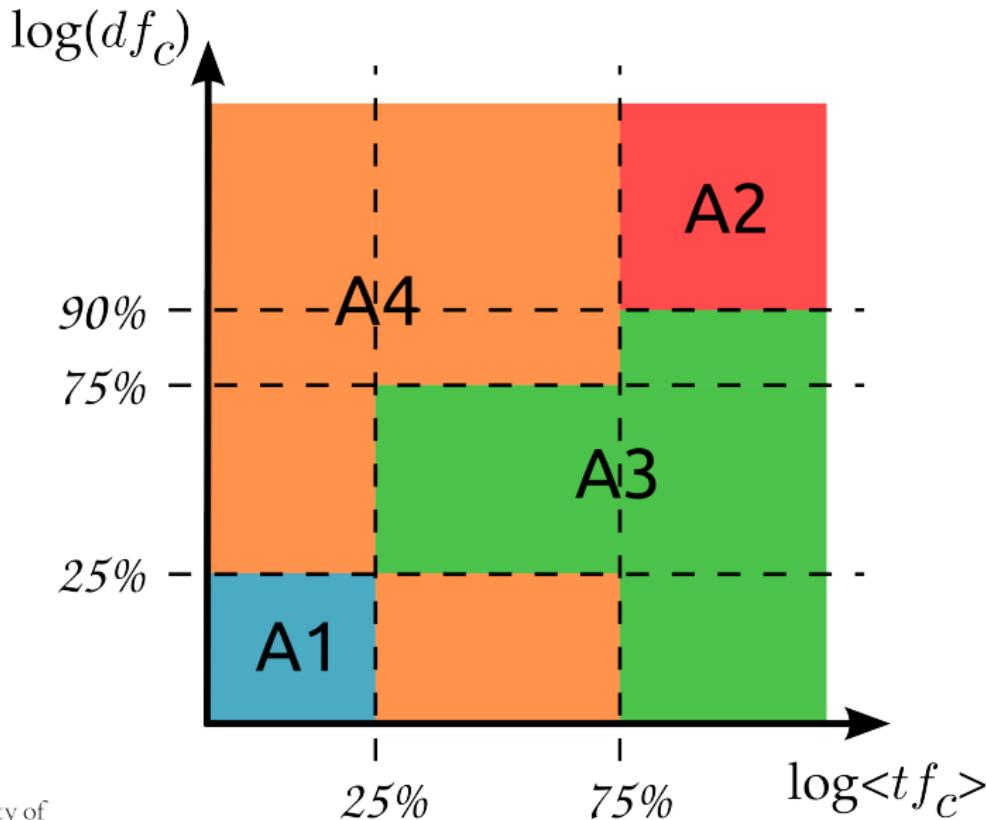
- average # of times a concept appears inside a document

$\langle tf_c \rangle \rightarrow$ average term frequency

- D. Jurafsky and J. Martin "Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition" Prentice Hall (2000).

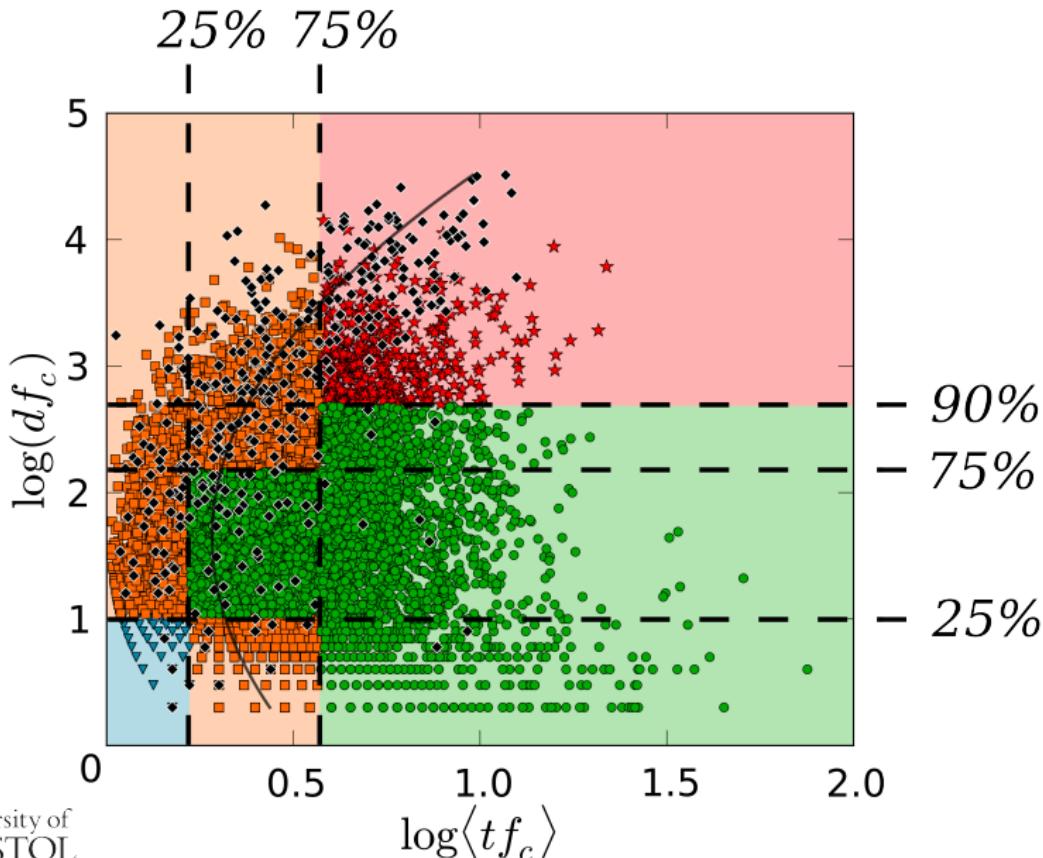


Bidimensional tessellation

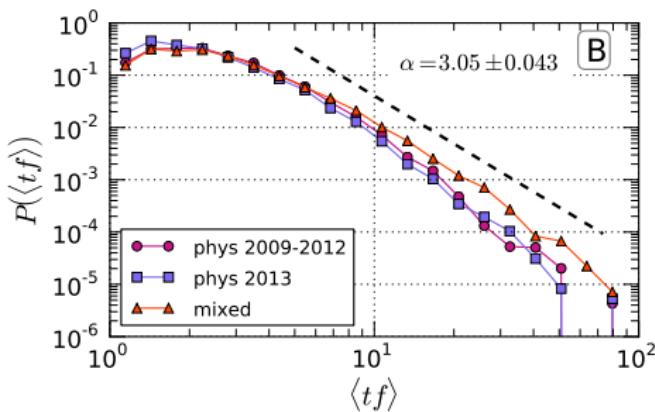
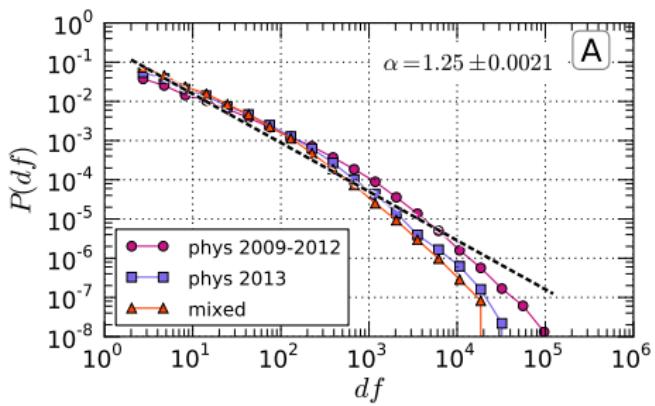


KUKUAEWm

Bidimensional tessellation



Bidimensional tessellation

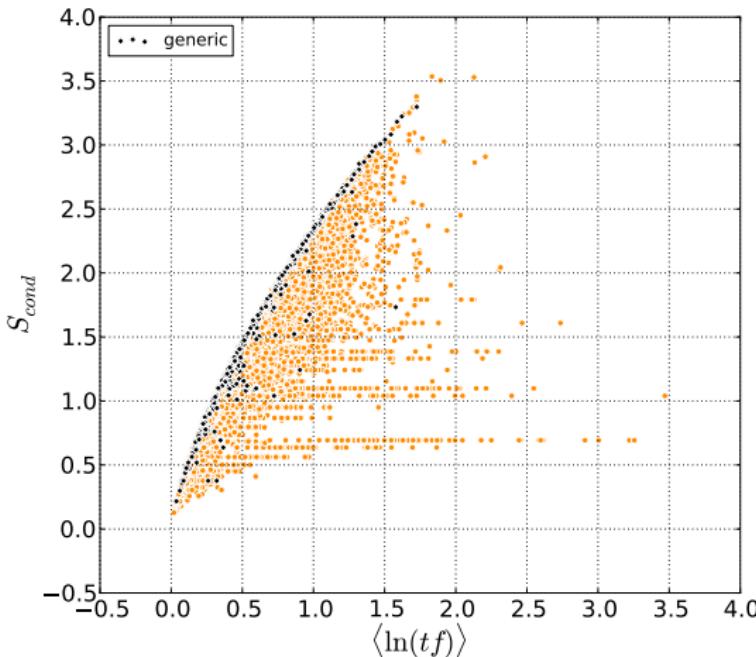


- F. Font-Clos *et al.* “A scaling law beyond Zipf’s law and its relation to Heaps’ law”. New J. Phys. **15** 093033 (2013).
- M. Gerlach *et al.* “Scaling laws and fluctuations in the statistics of . . .”. New J. Phys. **16** 113010 (2014).

Entropic filtering

KUKUAEWm

Maximum entropy



$$S(c) = - \sum_{j=1}^{\infty} p_c(j) \ln p_c(j)$$

- A. Berger et al. "A Maximum Entropy Approach to Natural Language . . .". Computational Linguistics **22** 39 (1996).

Maximum entropy

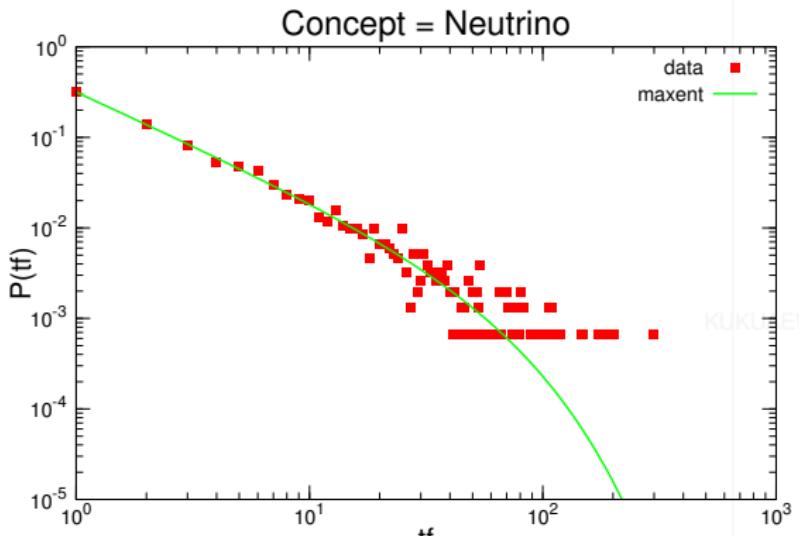
$$\sum_n p_n = 1$$

$$\sum_n p_n n = \langle n \rangle$$

$$\sum_n p_n \ln n = \langle \ln n \rangle$$

$$\ln p_n + \lambda n + \mu \ln n = 0$$

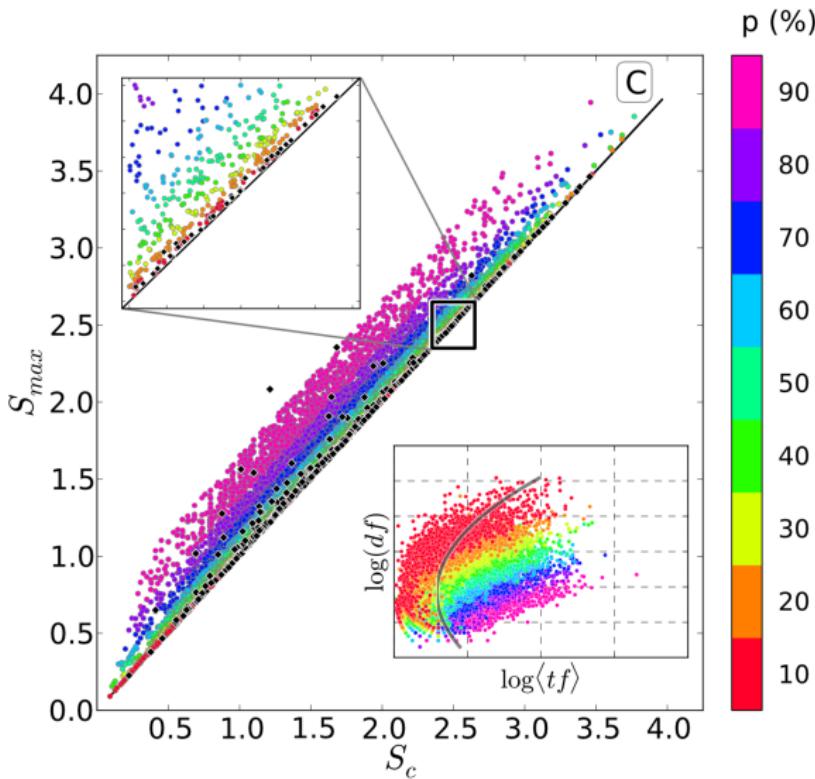
$$p_n = \frac{1}{Z} e^{-\lambda n} n^{-\mu}$$



- M. Gerlach *et al.* "Scaling laws and fluctuations in the statistics of . . .". New J. Phys. **16** 113010 (2014)
- R. Ferrer i Cancho *et al.* "Least effort and the origins of scaling in . . .". Proc. Nat. Acad. Sci. USA **100** 788 (2003).



Maximum entropy



Residual Entropy

$$S_d(c) = S_{max}(c) - S_c(c)$$

Results

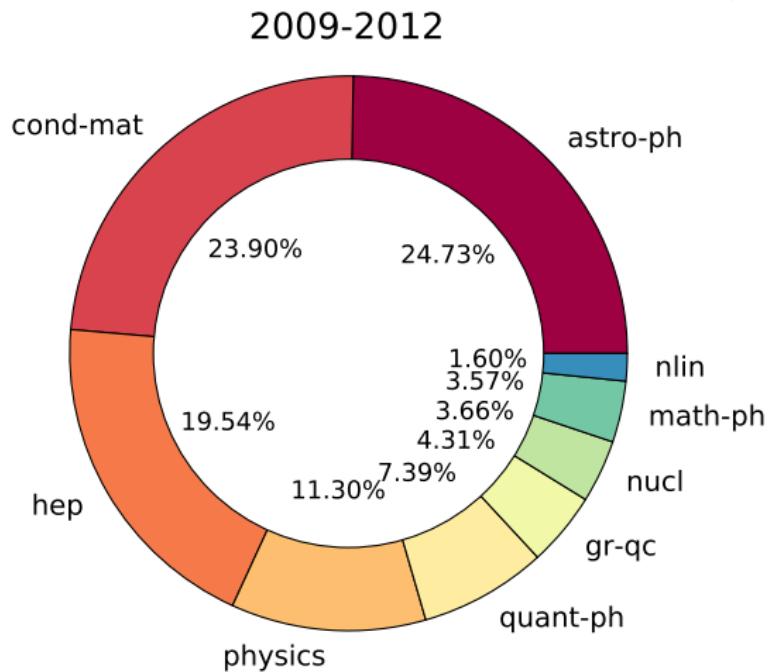
KUKUAEWm

Data

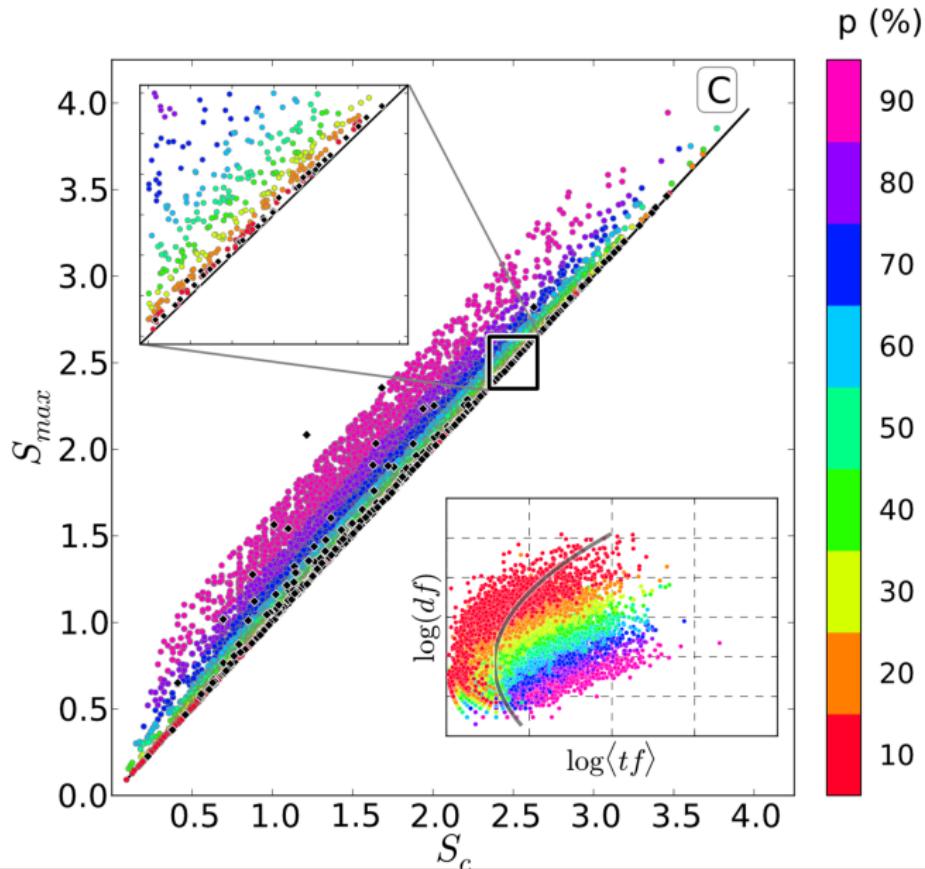
Relevant info.

arXiv Physics
2009 – 2012

N_a	189759
N_{con}	15040
T^*	10



What is a “generic concept”?



KUKUAEWm

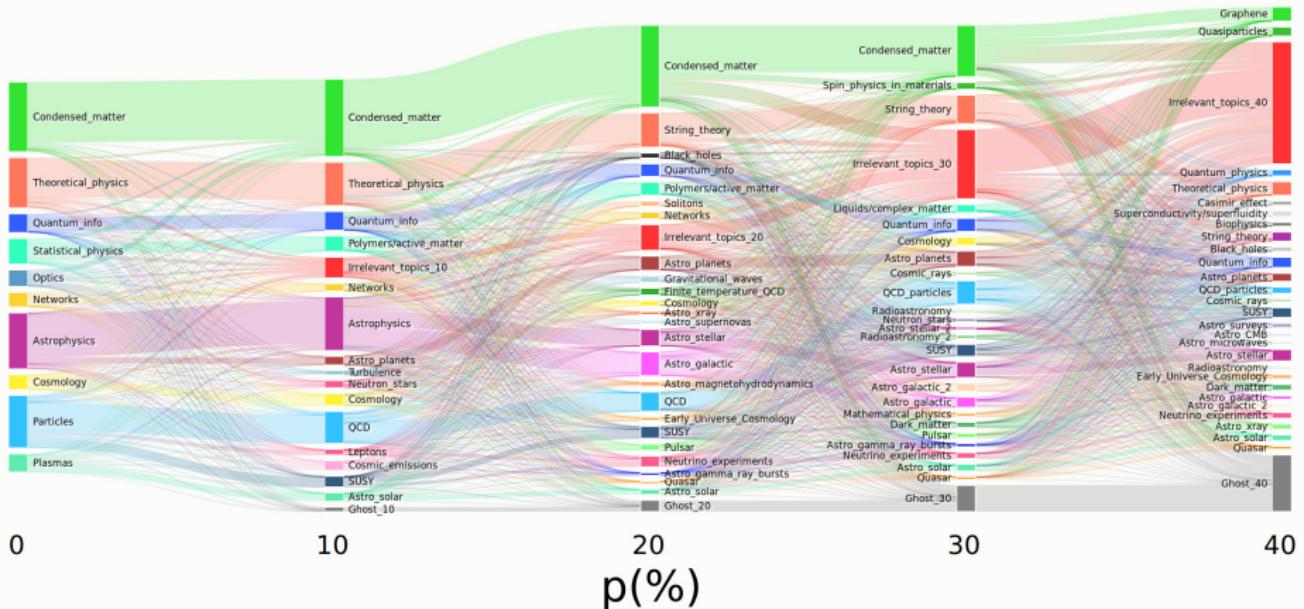
Effects of pruning concepts



▶ Link

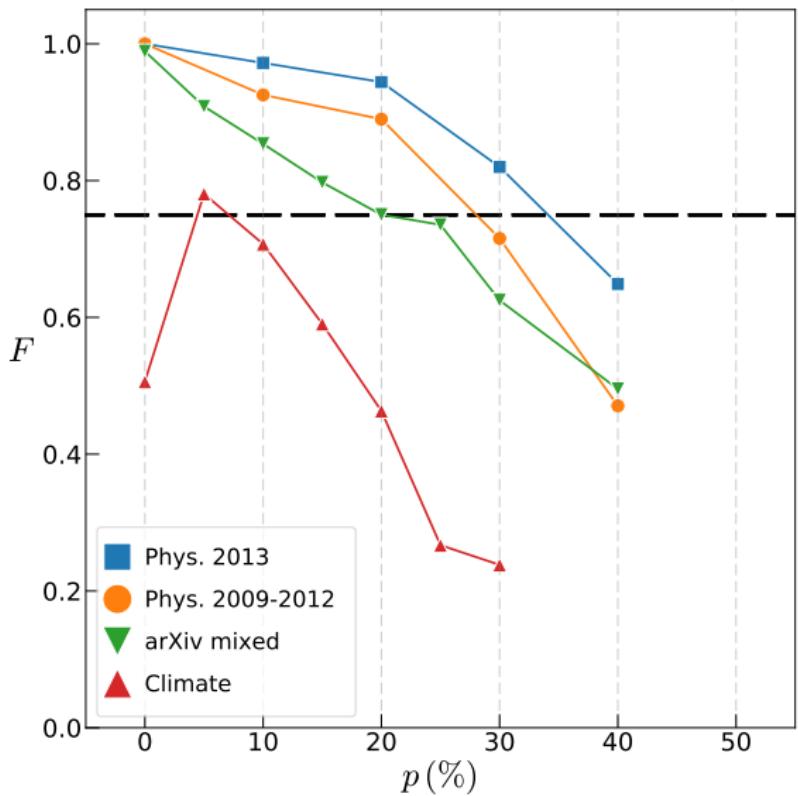


Effects of pruning concepts



Optimal filtering

$$F(p) = \frac{N_{good}(p)}{N_{tot}} =$$
$$= 1 - \frac{N_a^G(p) + N_a^I(p)}{N_{tot}}$$



Perspectives & Conclusions

KUKUAEWm

Perspective on Complex Systems

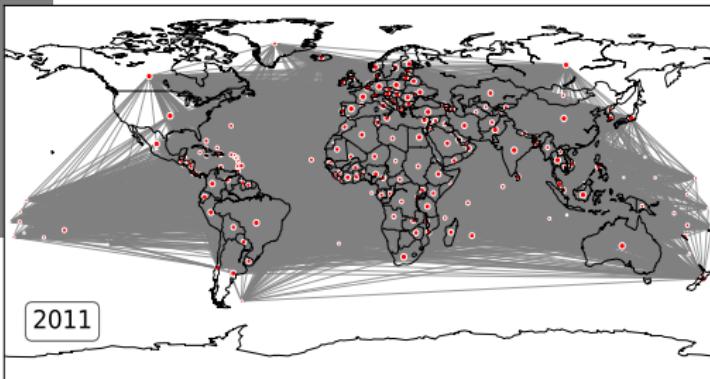
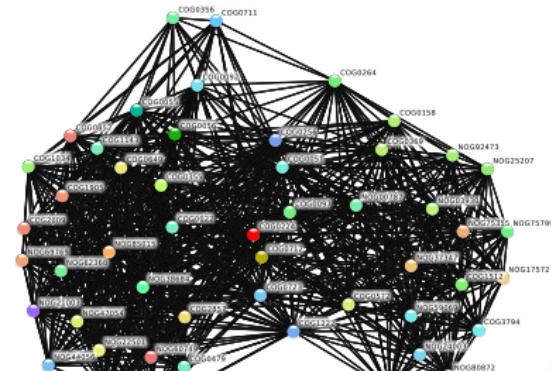
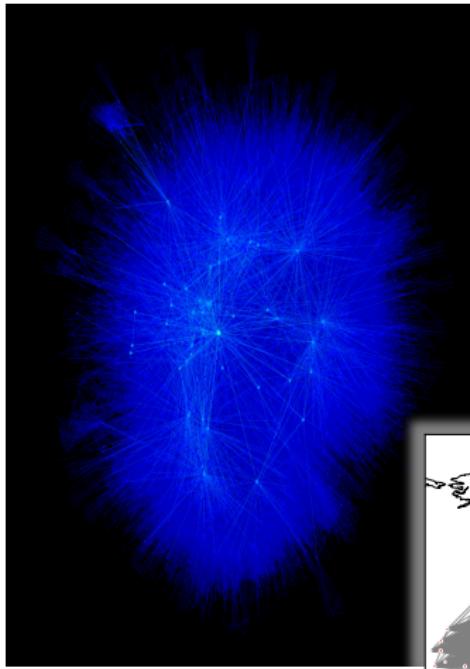


Using networks to study complex systems is like paleontology ...

Perspective on Complex Systems



Perspective on Complex Systems



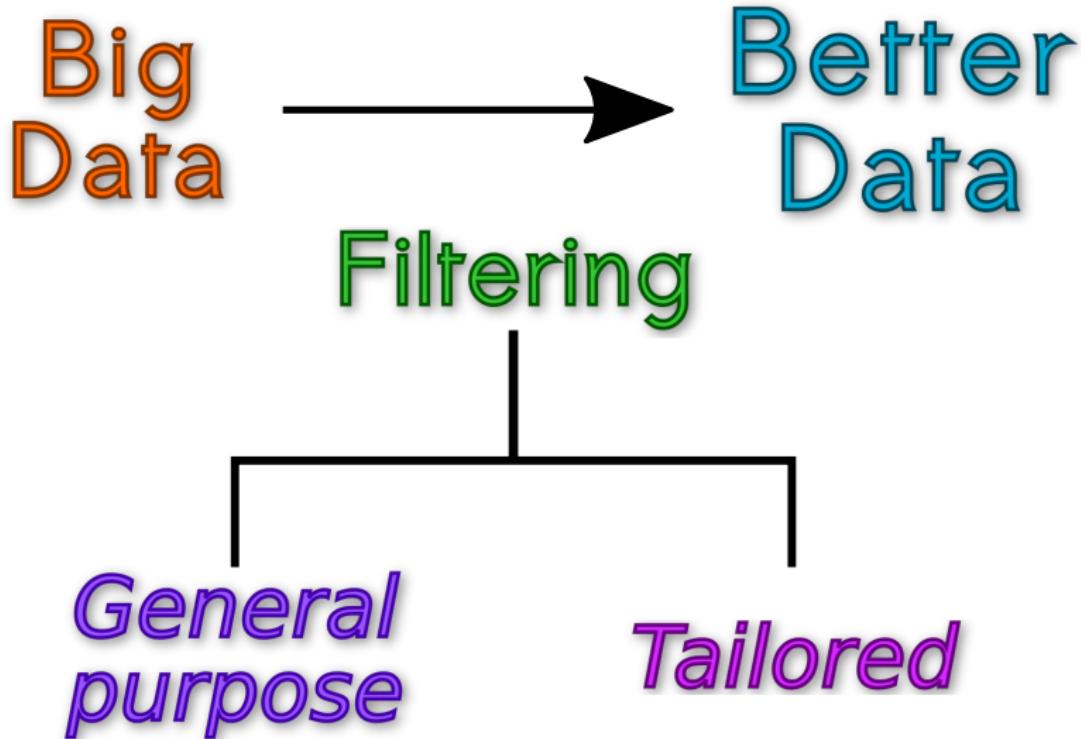
Perspective on Complex Systems

Question:

What can we **learn** from a complex system
whose **network representation** is **noisy** and/or
extremely **dense**?

IAEWm

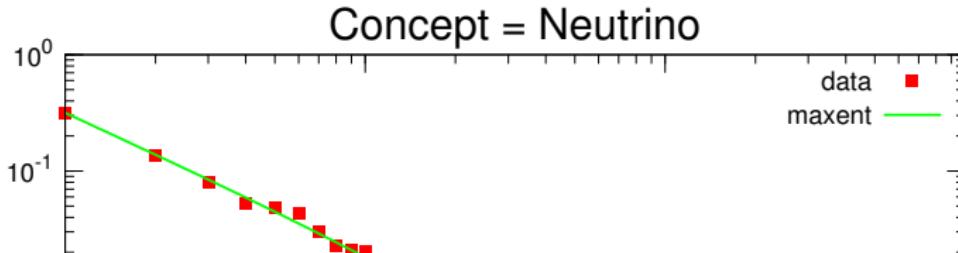
Perspective on Complex Systems



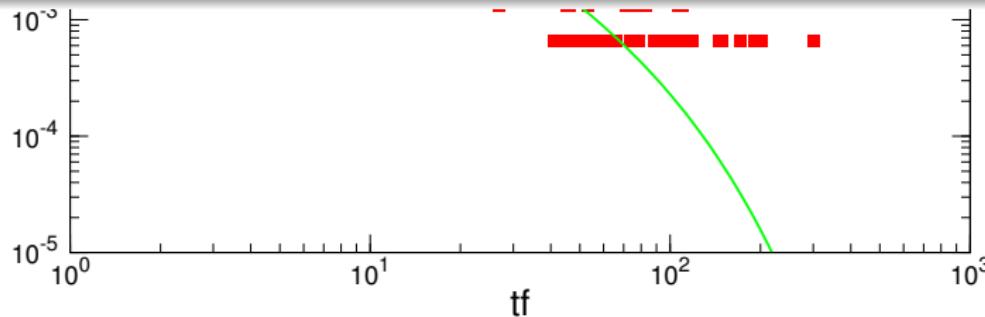
KUKUAEWm



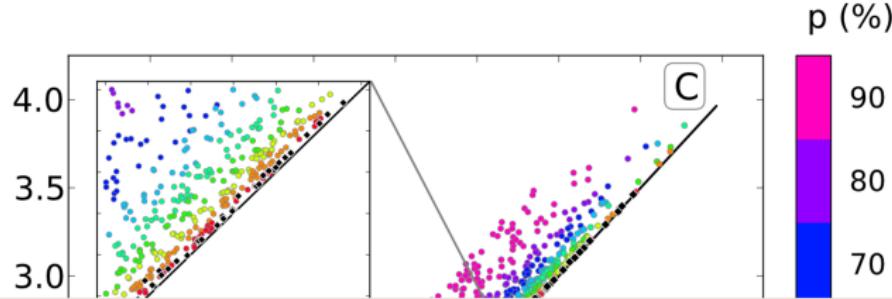
Summing up . . .



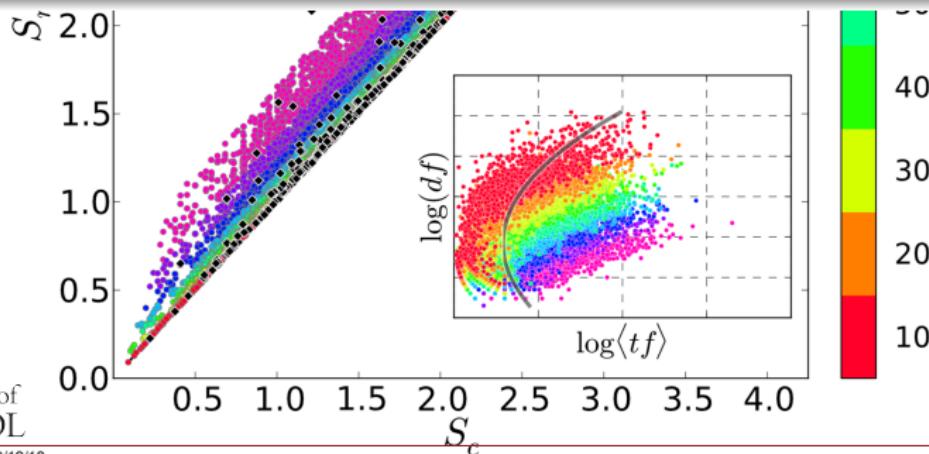
We have used the maximum entropy principle to build a method to prune out (**filter**) concepts used to extract topics.



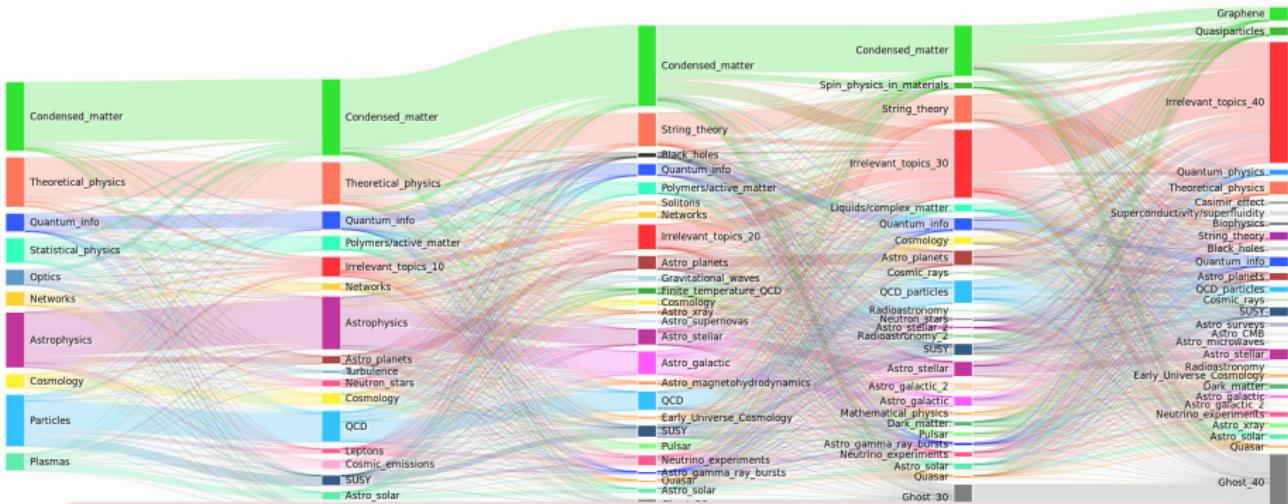
Summing up . . .



The method allows to identify collection dependent “**relevant concepts**” without requiring user validation.



Summing up . . .

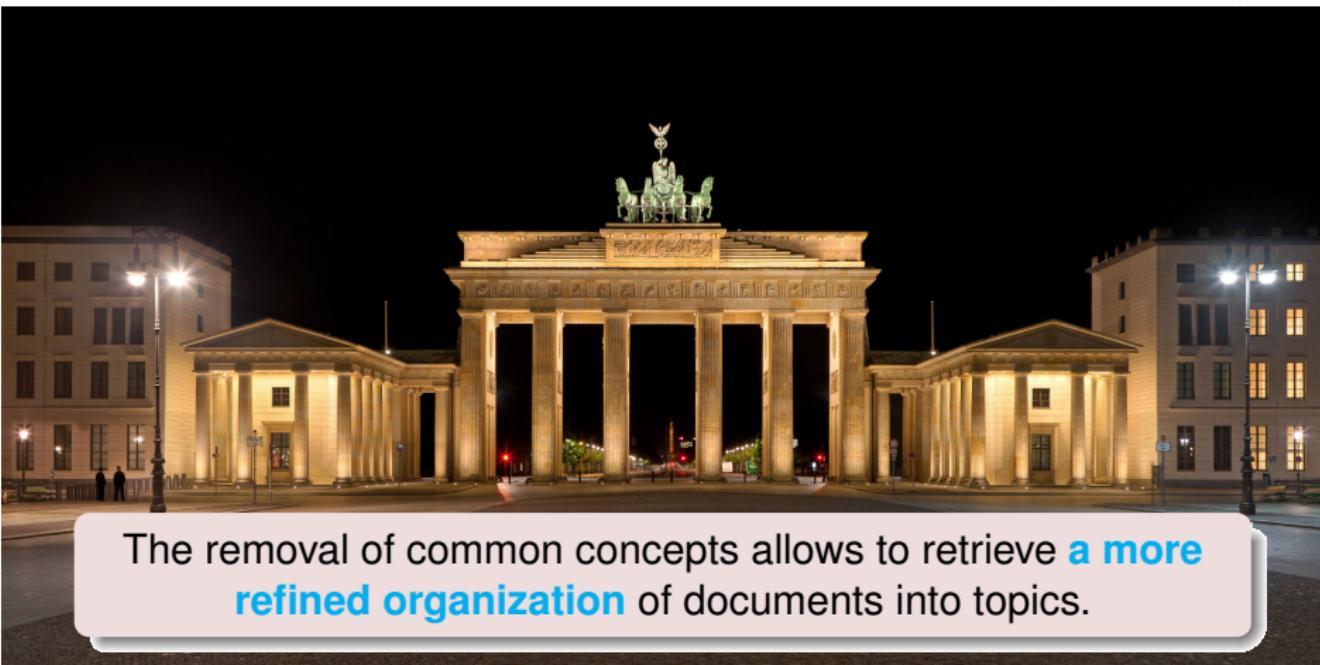


0

The removal of common concepts allows to retrieve a more refined organization of documents into topics.

40

Summing up . . .



The removal of common concepts allows to retrieve **a more refined organization** of documents into topics.

Acknowledgements



Paolo De Los Rios



Andrea Martini



Alex Constantin
&
ScienceWISE team

Acknowledgements



**FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION**

KUKUAEWm

Acknowledgements

arXiv.org > physics > arXiv:1705.06510

Physics > Physics and Society

Entropic selection of concepts unveils hidden topics in documents corpora

Andrea Martini, Alessio Cardillo, Paolo De Los Rios

(Submitted on 18 May 2017 (v1), last revised 11 May 2018 (this version, v2))

The organization and evolution of science has recently become itself an object of scientific quantitative investigation, thanks to the wealth of information contained in documents, such as citations between papers and co-authorship between researchers. However, only few studies have focused on the conceptual structure of the document corpus. In this paper we show how concepts can be extracted and analyzed, revealing the deeper organization of scientific knowledge. Unfortunately, several concepts can be so common that they do not give rise to any meaningful structure. To address this problem, we introduce a method to gauge the relevance of concepts according to the emergence of the underlying topical structure of the document corpus, because they give rise to a large amount of spurious and trivial relations. We apply our method to a collection of scientific documents and find that it is able to identify the most relevant concepts. By progressively removing concepts that, according to this metric, can be considered as generic, we find that the topic organization of the document corpus becomes more organized and structured.

alessio.cardillo@bristol.ac.uk

<http://www.bifi.es/~cardillo/>

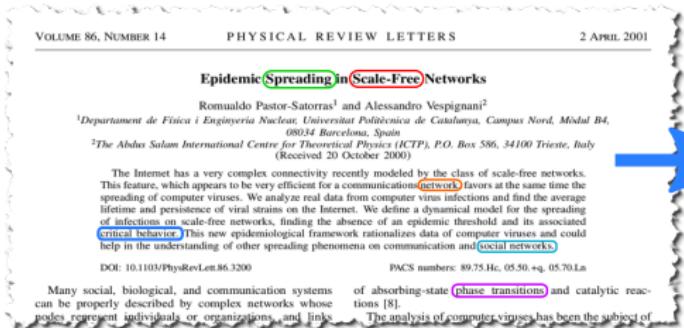
@a_cardillo

Part II

Appendix

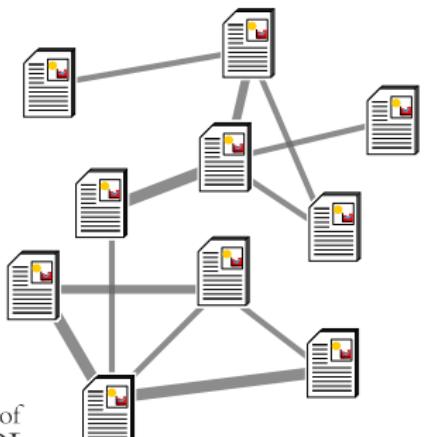
KUKUAEWm

Networks of documents

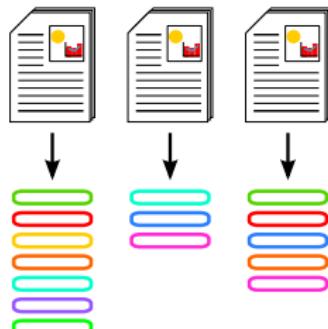


(a)

Scale-Free
Social network
Phase transition
Network
Spreading



(c)



(b)

TF-IDF and similarity

α	X	X	X	0	X	0	0
	C_1	C_2	C_3	C_4	C_5	C_6	C_7

KUKUAEWm

$$TF-IDF_{\alpha c} = u_{\alpha c} = \underbrace{tf_{\alpha c}}_{TF} \underbrace{\log \left(\frac{1}{df_c} \right)}_{IDF} = tf_{\alpha c} \log \left(\frac{N}{N_c} \right).$$

TF-IDF and similarity

Edge weight

$$w_{\alpha\beta} = \frac{\vec{u}_\alpha \cdot \vec{u}_\beta}{\|\vec{u}_\alpha\| \|\vec{u}_\beta\|},$$

$$w_{\alpha\beta} \in [0, 1],$$

β	2	43	0	18	0	11	27
α	13	5	9	0	30	0	0

$C_1 \quad C_2 \quad C_3 \quad C_4 \quad C_5 \quad C_6 \quad C_7$

TF-IDF and similarity

Edge weight

$$w_{\alpha\beta} = \frac{\vec{u}_\alpha \cdot \vec{u}_\beta}{\|\vec{u}_\alpha\| \|\vec{u}_\beta\|},$$

$$w_{\alpha\beta} \in [0, 1],$$

$$\begin{aligned} w_{\alpha\beta} &= \frac{(13 \times 2) + (43 \times 5)}{55.02 \times 34.28} = \\ &= \frac{241}{1886.09} \simeq 0.13. \end{aligned}$$

β	2	43	0	18	0	11	27
α	13	5	9	0	30	0	0
	C_1	C_2	C_3	C_4	C_5	C_6	C_7

All collections

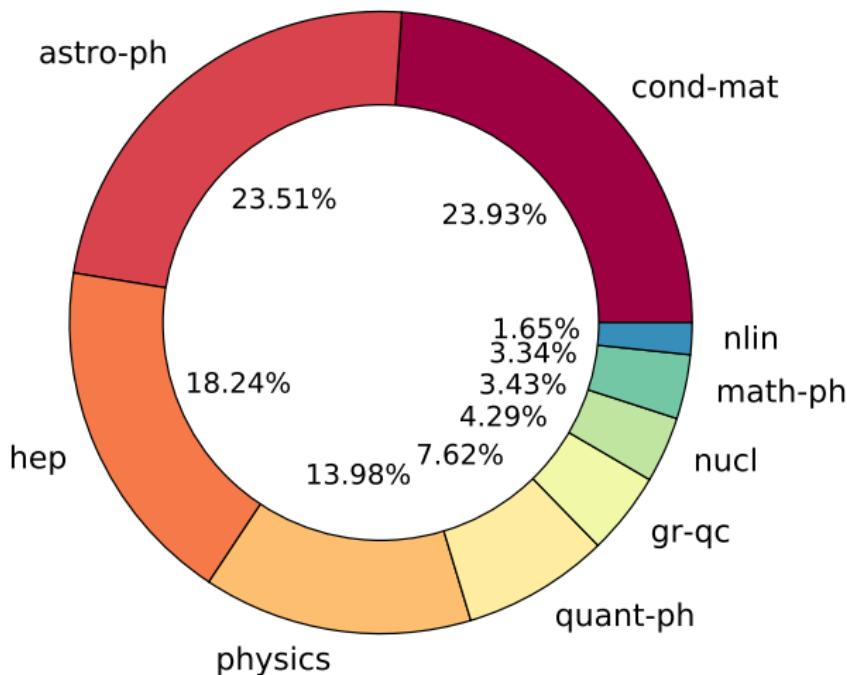
Documents collections

	N_{con}	N_a	T	T^*	$\langle N_a \rangle_{T^*}$	$\langle N_{con} \rangle_{T^*}$
arXiv Physics 2013	13173	52979	10	10	5298	4212
arXiv Physics 2009 – 2012	15040	189759	10	10	18976	6185
arXiv Mixed	19843	50408	14	12	4155	3994
Climate change webdocs	822545	18770	201	22	432	26004

IAEWm

arXiv Collections

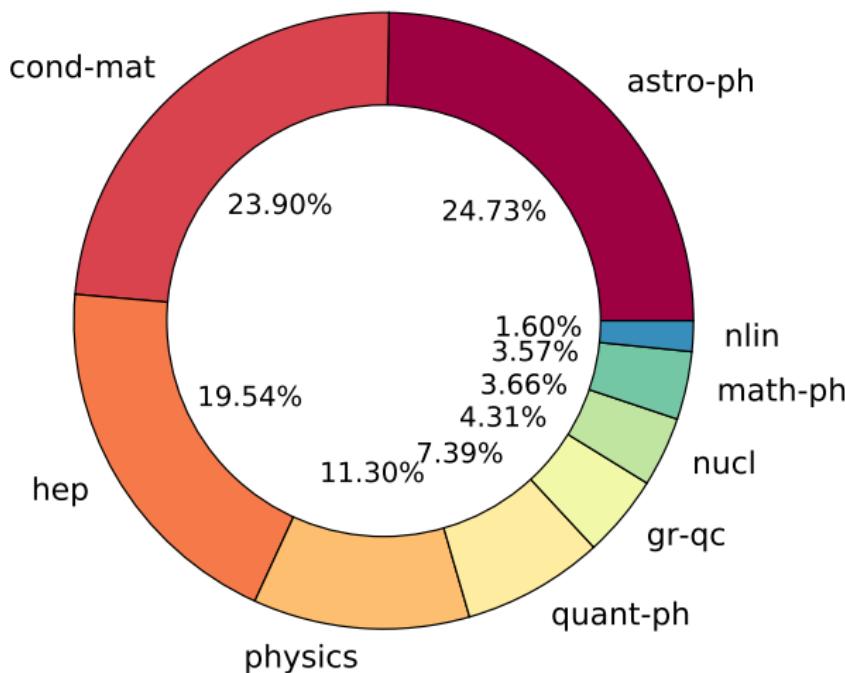
2013



KUKUAEWm

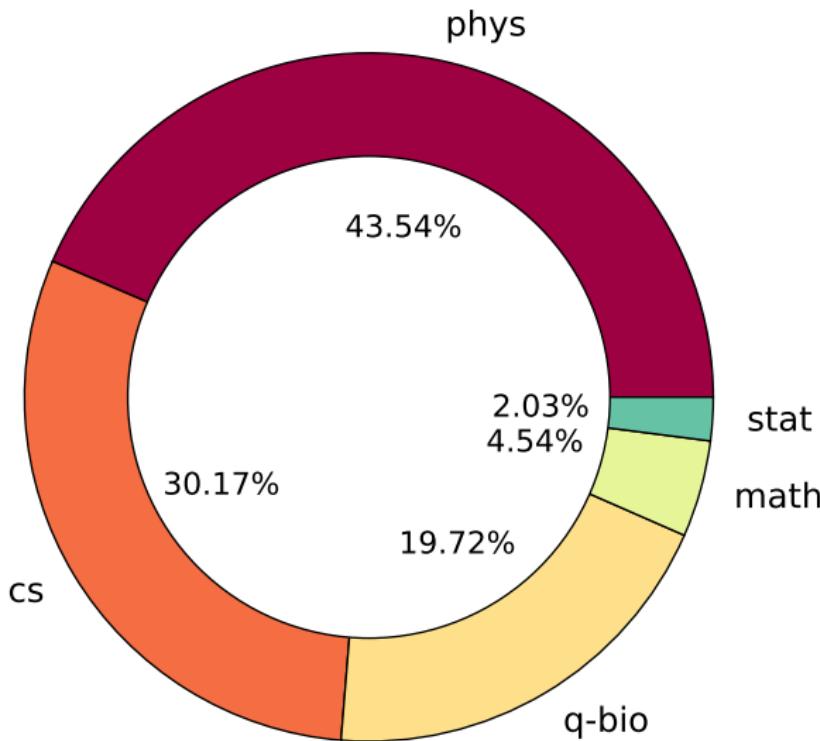
arXiv Collections

2009-2012



KUKUAEWm

arXiv Collections

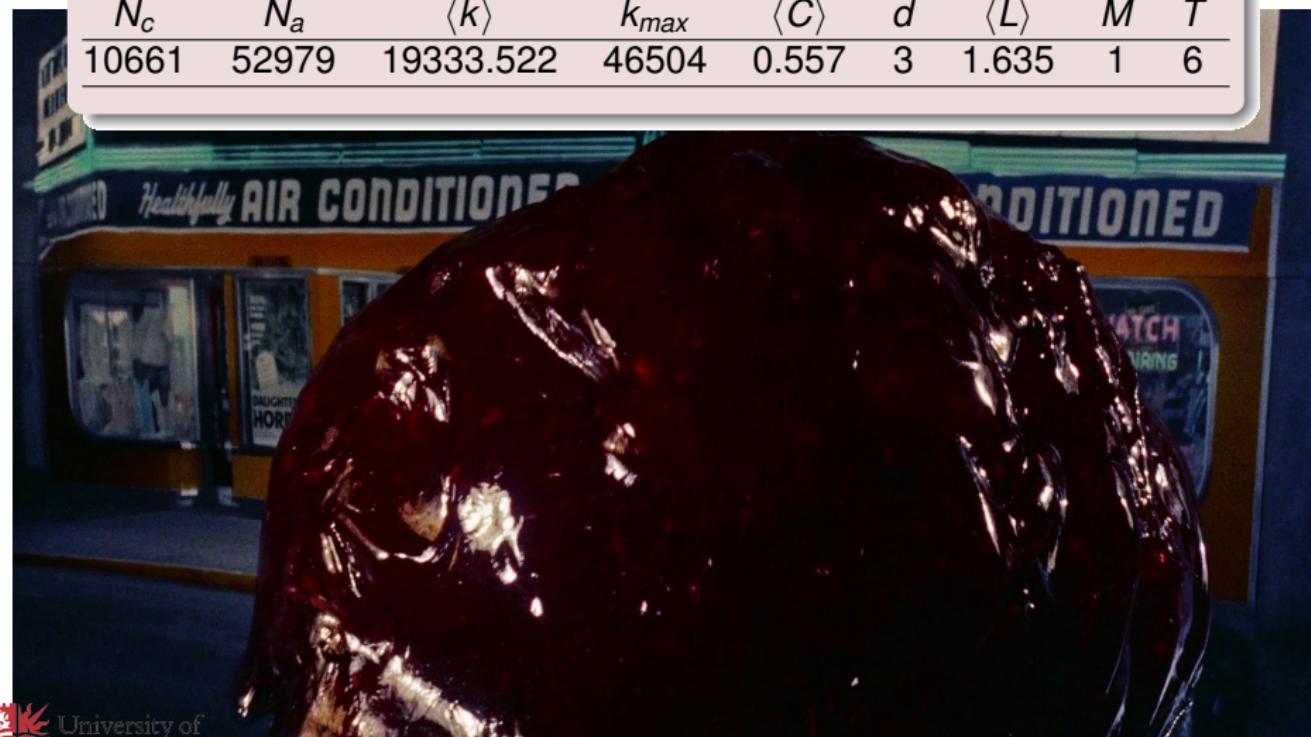


KUKUAEWm

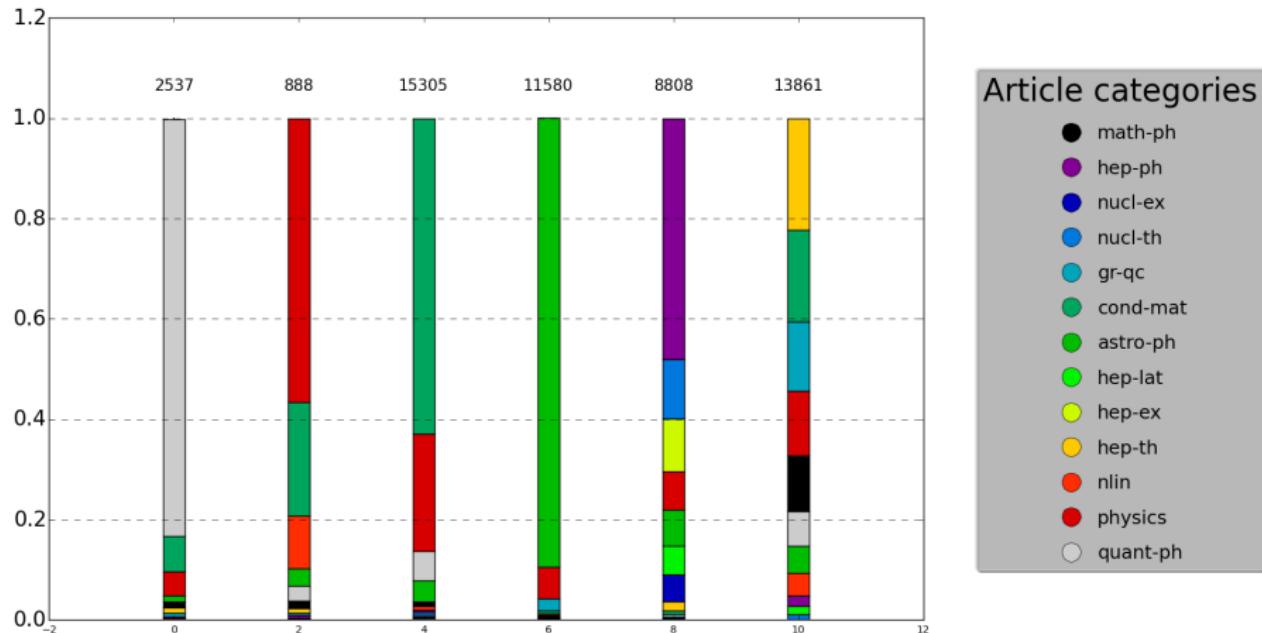
The data: 2013 Physics arXiv

Network properties

N_c	N_a	$\langle k \rangle$	k_{max}	$\langle C \rangle$	d	$\langle L \rangle$	M	T
10661	52979	19333.522	46504	0.557	3	1.635	1	6

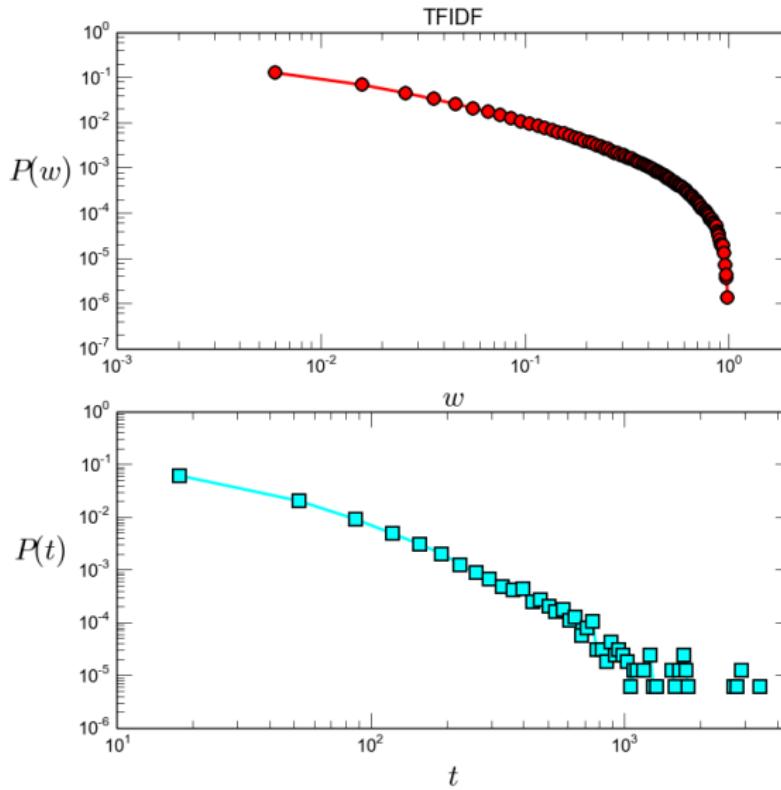


The data: 2013 Physics arXiv



$$\rho = \frac{K}{K_{\max}} \simeq 36\%$$

Edge pruning/sparsification methods



KUKUAEWm

Edge pruning/sparsification methods

Institution: EPFL

Proceedings of the National Academy of Sciences of the United States of America

CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // FOR AUTHORS // ABOUT PNAS // COLLECTED ARTICLES // BROWSE BY TOPIC // EARLY EDITION

Home > Current Issue > vol. 106 no. 16 > M. Ángeles Serrano, 6483–6488, doi: 10.1073/pnas.0808904106

 CrossMark
click for updates

Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano^{a,1}, Marián Boguña^b and Alessandro Vespignani^{c,d}

Author Affiliations

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved March 2, 2009 (received for review September 9, 2008)

Abstract | Full Text | Authors & Info | Figures | SI | Metrics | Related Content |  |  +SI

Abstract

This Issue

PNAS April 21, 2009
vol. 106 no. 16
Masthead (PDF)
Table of Contents

◀ PREV ARTICLE ▶ NEXT ARTICLE ▶

 View this article with LENS beta

Don't Miss

- Serrano M.A., et al. *Extracting the multiscale backbone of complex weighted networks*. Proc. Natl. Acad. Sci. (USA) 106 6483 (2009).



Edge pruning/sparsification methods

PHYSICAL REVIEW E

statistical, nonlinear, and soft matter physics

Highlights Recent Accepted Authors Referees Search About 

Information filtering in complex weighted networks

Filippo Radicchi, José J. Ramasco, and Santo Fortunato
Phys. Rev. E **83**, 046101 – Published 1 April 2011

Article

References

Citing Articles (8)

PDF

HTML

Export Citation



ABSTRACT

Many systems in nature, society, and technology can be described as networks, where the vertices are the system's elements, and edges between vertices indicate the interactions between the corresponding elements. Edges may be weighted if the interaction strength is measurable. However, the full network information is often redundant because tools and techniques from network analysis

- Radicchi, F., et al. *Information filtering in complex weighted networks*. Physical Review E, **83** 046101. (2011).

Edge pruning/sparsification methods

arXiv.org > physics > arXiv:1706.00230

Physics > Physics and Society

Irreducible network backbones: unbiased graph filtering via maximum entropy

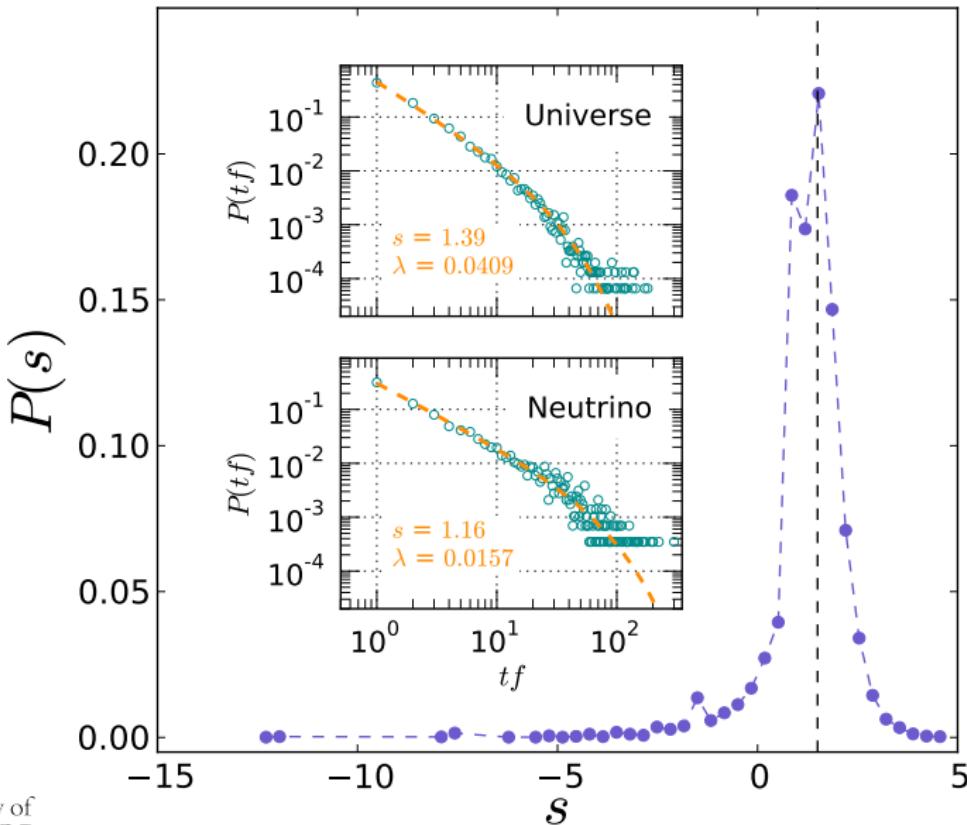
Valerio Gemmetto, Alessio Cardillo, Diego Garlaschelli

(Submitted on 1 Jun 2017 (v1), last revised 9 Jun 2017 (this version, v2))

Networks provide an informative, yet non-redundant description of complex systems only if links represent truly dyadic relationships that cannot be properties such as size, importance, or coordinates in some embedding space. In any real-world network, some links may be reducible, and others dichotomy persists despite the steady increase in data availability and resolution, which actually determines an even stronger need for filtering tech from non-essential ones. Here we introduce a rigorous method that, for any desired level of statistical significance, outputs the network backbone t nodes. i.e. their dearees and strengths. Unlike previous approaches. our method employs an exact maximum-entropy formulation guaranteeing tha

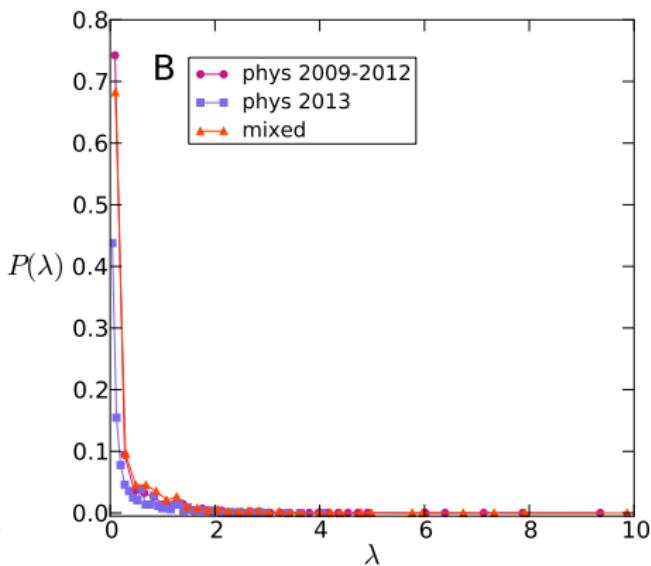
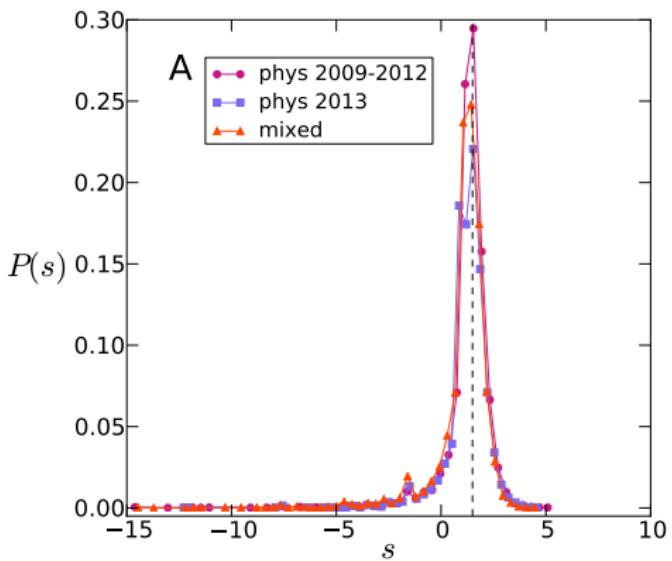
- Gemmetto, V., et al. *Irreducible network backbones: unbiased graph filtering via maximum entropy.* arXiv:1706.00230 (2017).

Maximum entropy – why power law?

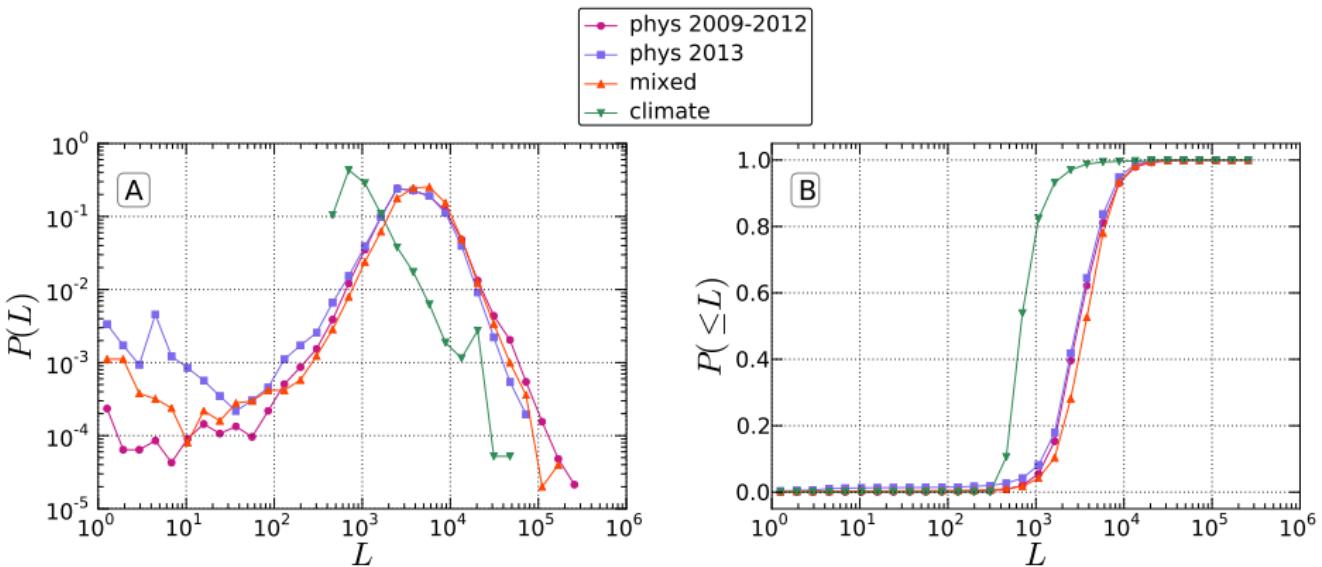


KUKUAEWm

Maximum entropy – why power law?



Maximum entropy – TF density

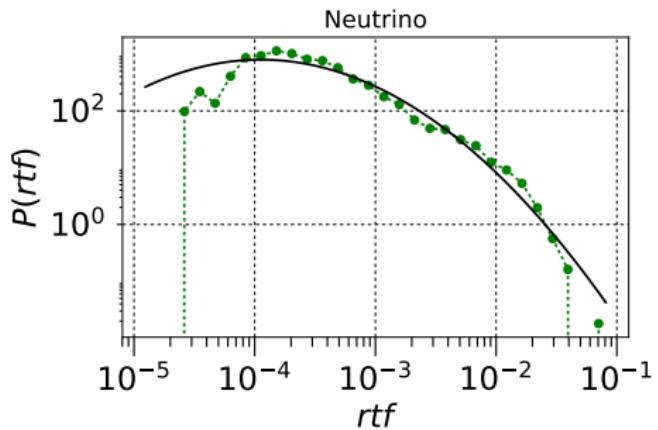


Maximum entropy – TF density

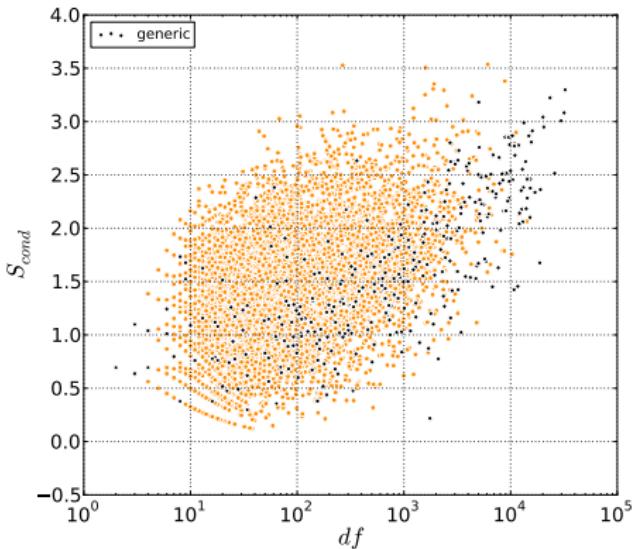
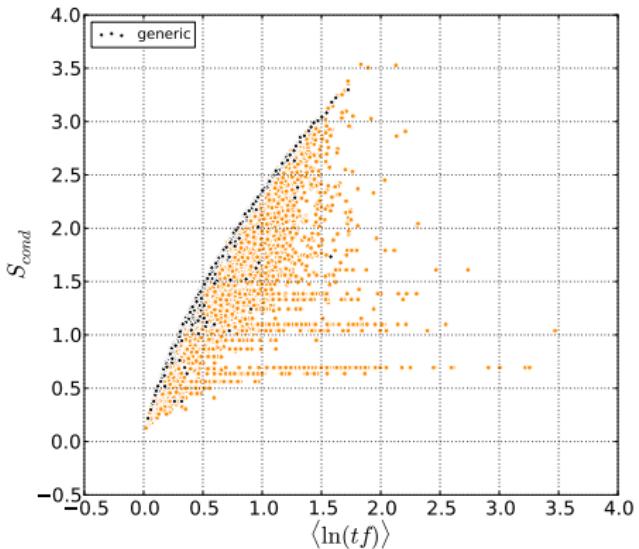
$$\sum_n p_n = 1$$

$$\sum_n p_n \ln n = \langle \ln n \rangle$$

$$\sqrt{\sum_n p_n (\ln n - \langle \ln n \rangle)^2} = \sigma_{\ln n}$$



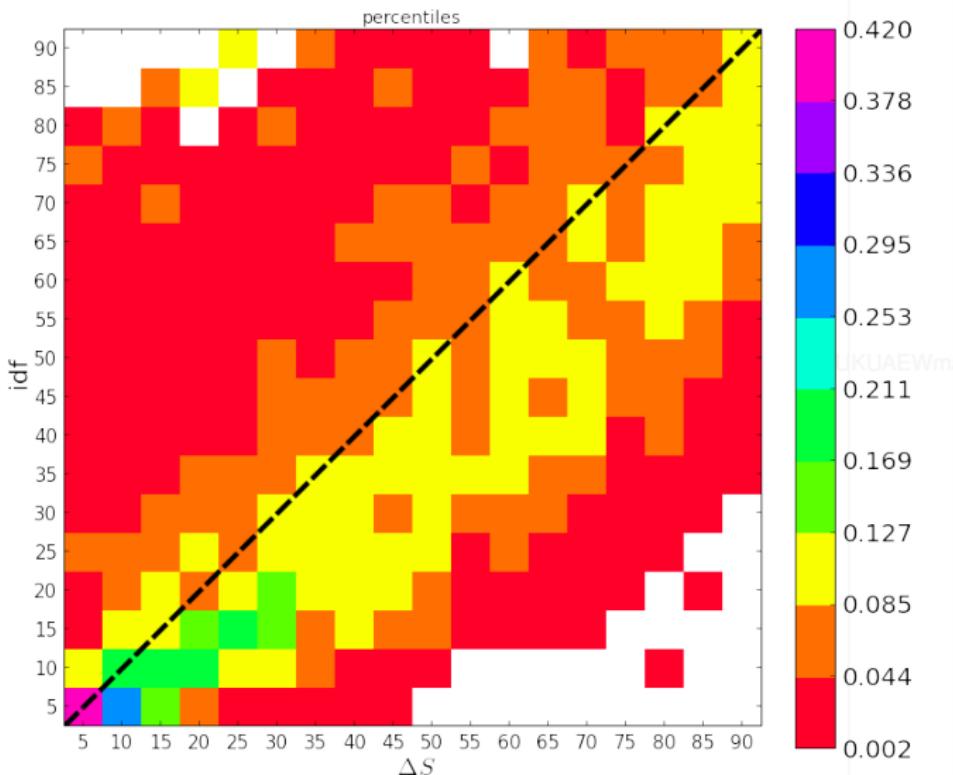
Why not df ?



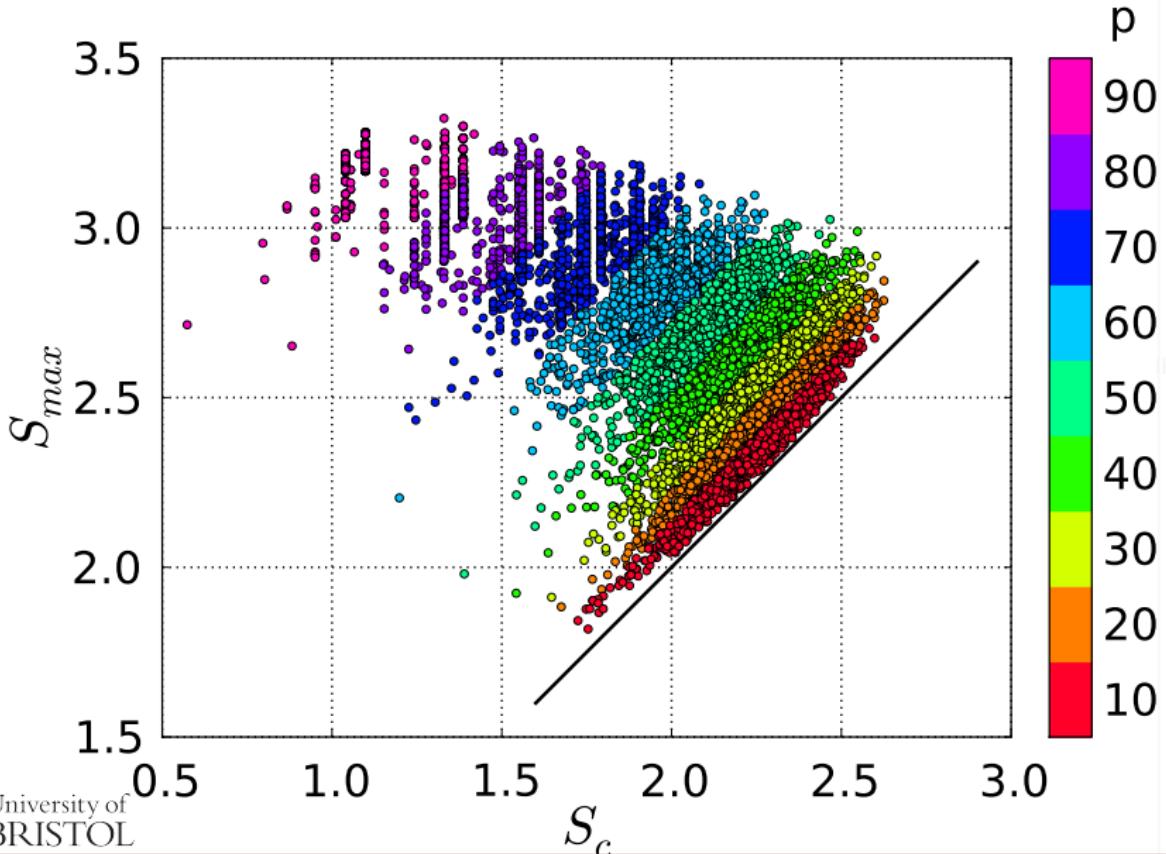
Why not df ?

Jaccard Score

$$J = \frac{|A \cap B|}{|A \cup B|}.$$



Climate dataset



Climate dataset

